Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables

Dustmann, C.; van Soest, A.H.O.

Publication date:
1999

Contents

Parametric and Semiparametric Estimation in Models with Misclassified Categorical Dependent Variables

Christian Dustmann[†] Arthur van Soest [‡]

May 1999

Abstract

We consider both a parametric and a semiparametric method to account for classification errors on the dependent variable in an ordered response model. The methods are applied to the analysis of self–reported speaking fluency of male immigrants in Germany. We find some substantial differences in parameter estimates between parametric and semiparametric models, and between predictions of true and reported fluency. We compare the predictions of the three models, and perform a graphical test of the parametric models against semiparametric alternatives.

key words: immigrants, speaking fluency, measurement error

JEL codes: C14,C35,J15

*We are grateful to Joel Horowitz, Marlene Mueller, Bertrand Melenberg and Marcel Das for providing Gauss programs and useful comments.

[†]University College London, and Institute for Fiscal Studies, London, e-mail: c.dustmann@ucl.ac.uk

[‡]Tilburg University, Department of Econometrics, P.O. Box 90153, 5000 LE Tilburg, Netherlands. e-mail: avas@kub.nl.

1 Introduction

Many empirical studies in economics and other social sciences are concerned with the analysis of ordered categorical dependent variables. Categorical data can be affected by misclassification error. This is especially the case if the categorical assignment is based on subjective self-reported evaluations, as used in many empirical analyses. Examples are studies which analyse data on job satisfaction (see, for example, Clark and Oswald (1994)), satisfaction with health (Kerkhofs and Lindeboom (1995)), or future expectations of household income (Das and Van Soest (1997)).

In applied work, nonlinear parametric limited dependent variable models are typically used for analysing categorical dependent variables. Misclassification can in this case lead to seriously biased parameter estimates, even if the parametric model correctly specifies the unobservable "true" categorical variable. To deal with this problem, estimators have been proposed for parametric binary choice models which correct for misclassification by explicitly incorporating the misclassification probabilities as additional parameters. Lee and Porter (1984) estimate an exogenous switching regression model for market prices of grain, distinguishing regimes where firms are cooperative and noncooperative. They observe an imperfect indicator of the actual regime, and estimate the two misclassification probabilities that one regime is observed given that the other regime is active. They then use these probabilities to correct the estimates of the price equations in each regime for the misclassification errors. Hausman et al. (1998) estimate binary choice models for job change. In their parametric models, they find significant probabilities of misclassifying in both directions. They also estimate a semiparametric model, and find that the semiparametric estimates are similar to the parametric estimates allowing for misclassification.

In this paper, we consider parametric and semiparametric models for ordered categorical dependent variables with more than two outcomes. Our parametric model generalizes the binary response models by Lee and Porter (1984) and Hausman et al. (1998) by incorporating probabilities of misclassifying outcomes into more than two

categories. Other than in the binary case, the identification of the misclassification probabilities for the intermediate categories relies on distributional assumptions. In this case, parametric estimates of the effects of the 'true' outcome may be sensitive to these distributional assumptions, so that semiparametric estimation can be particularly useful. We show that, if the misclassification probabilities are not too large, the parametric model is a special case of a single index model satisfying a weak monotonicity condition. This model can be estimated using the semiparametric technique of Horowitz and Haerdle (1996), combining average derivatives estimation with a GMM type of estimator to take account of discrete regressors.

We apply both parametric and semiparametric estimators to data on self–reported speaking fluency of male immigrants to Germany. A growing literature is concerned with the determinants of language fluency, and its effects on various economic activities (see, for instance, McManus, Gould and Welch (1983), Rivera-Batiz (1990), Chiswick (1991), and Chiswick and Miller (1995)). Nearly all studies are based on self-reported language categorisations. It is likely that this type of data is even more affected by misclassification than objective variables such as the job change variable in Hausman et al. (1998). Besides errors resulting from, for instance, misunderstanding of survey questions, responses of this type may be misclassified because of heterogeneity in the underlying subjective standards. Dustmann and van Soest (1998) provide some evidence that misclassification is substantial in this data. They compare responses of the same individuals at different points in time. They find that, under the assumption that a deterioration of language capacity is not possible, one fourth of the total variance in the language indicator is due to misclassification. This number is a lower bound, since it accounts only for errors which are not time persistent.

The results of our analysis show that allowing for misclassification errors has some effect on the estimates of the parameters in the speaking fluency equation. The parametric model which allows for misclassification is a clear improvement to the standard ordered probit model. The estimated probabilities of misclassification into the extreme

categories are large. A comparison of the twoparametric models shows that the predictions for the observed outcome variable and for the true outcome variable differ considerably. The misclassification model leads to predicted probabilities closer to zero or one. A formal test of the parametric models against a semiparametric alternative is proposed, based upon uniform confidence bands of the nonparametric regression function of the dependent variable on the parametrically estimated index. For both parametric models, we find that the parametric regression function is contained in the confidence bands, so that the parametric models cannot be rejected.

The paper is organized as follows. In section 2, we present the models and their estimators. In section 3.1, we briefly describe the data for our empirical application. Parametric and semiparametric estimates are presented in Section 3.2. In Section 3.3, we compare predictions of the two parametric models and the semiparametric model, and discuss some specification tests of the parametric models. Section 4 concludes.

2 Categorical Data and Misclassification

For simplicity, we assume that the dependent variable is observed on an ordinal scale with three levels, coded 1, 2 and 3. All results extend straightforwardly to the case of more than three categories. Starting point is the standard ordered probit model, not allowing for misclassification errors. It assumes observed categorical information is related to an underlying latent index y^* as follows (the index indicating the individual is suppressed):

$$y^* = x'\beta + u, \tag{1}$$

$$y = j \quad \text{if} \quad m_{j-1} < y^* < m_j, \quad j = 1,2,3\,, \tag{2}$$

$$u|x \sim N(0,\sigma^2)\,. \tag{3}$$

Here x is a vector of explanatory variables including a constant term, β is the vector of parameters of interest, and u is the error term. We assume $m_0 = -\infty$,

$m_1 = 0, m_3 = \infty$. The variance σ^2 and the bound m_2 can be seen as nuisance parameters. A normalization of the scale has to be added for identification. This will be discussed below. Throughout, we assume that the (y, x) are a random sample from the population of interest.

2.1 A Parametric Misclassification Model

For the binary choice case, Hausman et al. (1998) show that the bias in estimates of β may be substantial if some observations on the endogenous variable suffer from misclassification. They propose a generalization of the binary probit model to take account of misclassification errors. We extend their framework for the binary probit model to ordered probit.

Assume that the individual is observed to be in category y, but that the (unobserved) true category is z, which is related to the latent variable y^* as in the ordered probit case:

$$z = j \quad \text{if} \quad m_{j-1} < y^* < m_j\,, \quad j = 1, 2, 3\,. \tag{4}$$

The probabilities of misclassification are given by:

$$\text{Prob}(y = j | z = k) = p_{k,j}, \quad j, k = 1, 2, 3, j \neq k, \tag{5}$$

where $p_{k,j}$ is the probability that an observation which belongs in category k is classified in category j. If $p_{k,j} = 0$ for all j, k with $j \neq k$, then there is no misclassification and the model simplifies to the ordered probit model.

In a model with three categories, there are six misclassification probabilities $p_{k,j}$. Thus compared to the standard ordered probit for three categories, this model has six additional parameters.

For the binary choice case (with categories denoted 0 and 1), Hausman et al. (1998) show that identification of $p_{k,j}$, $j, k = 0, 1$ does not rely on the normality assumption, as long the support of $x'\beta$ is the whole real line, i.e. as long as observations with very

low and very high values of $x'\beta$ occur with nonzero probability. The probabilities of misclassification are then given by:

$$p_{1,0} = \lim_{x'\beta \to \infty} \mathrm{P}(y=0|x) \text{ and } p_{0,1} = \lim_{x'\beta \to -\infty} \mathrm{P}(y=1|x)\,.$$

If $p_{0,1}$ and $p_{1,0}$ do not depend on x and if u is independent of x, the model satisfies the single index property: $E\{y|x\}$ depends on x via $x'\beta$ only. Therefore, β is identified up to scale and sign. The additional condition required for identification is that $p_{0,1}$ and $p_{1,0}$ are not too large:

$$p_{1,0} + p_{0,1} < 1. \tag{6}$$

This condition guarantees that $E\{y|x\}$ increases with $x'\beta$. Accordingly, the sign of β is also identified, and it then follows from (5) that the $p_{0,1}$ and $p_{1,0}$ are nonparametrically identified.

For the ordered probit case with three categories coded 1, 2 and 3, and with misclassification probabilities, we obtain

$$\begin{aligned} E\{y|x\} &= 2 - p_{2,1} + p_{2,3} - \Phi((m_1 - x'\beta)/\sigma)(1 - p_{1,2} - p_{2,1} + p_{2,3} - 2p_{1,3}) \\ &+ [1 - \Phi((m_2 - x'\beta)/\sigma)](1 - p_{3,2} - p_{2,3} + p_{2,1} - 2p_{3,1})\,. \end{aligned} \tag{7}$$

Thus the condition that $E\{y|x\}$ increases with $x'\beta$ for every value of $x'\beta$ implies, instead of (6) for the binary choice case,

$$p_{1,2} + p_{2,1} - p_{2,3} + 2p_{1,3} < 1 \text{ and } p_{2,3} + p_{3,2} - p_{2,1} + 2p_{3,1} < 1\,. \tag{8}$$

This condition is satisfied for small enough values of the misclassification probabilities.

The argument for nonparametric identification in the binary choice case applies to $p_{1,j}$ and $p_{3,j}$, but not to $p_{2,1}$ or $p_{2,3}$. Identification of these is achieved in this parametric model by imposing normality of the error terms. The model can straightforwardly be estimated by Maximum Likelihood (ML), where the $p_{k,j}$ are estimated jointly with

the slope parameters β. The ML estimates are consistent, asymptotically normal, and asymptotically efficient if the assumptions (including normality of the errors) are satisfied. They will generally be inconsistent if the errors are not normally distributed.

2.2 A Semiparametric Approach

The parametric ML estimates of the slope parameters β require distributional assumptions and may not be robust to misspecification. If we are interested in β only and consider the $p_{k,j}$ as nuisance parameters, semiparametric estimation seems a good alternative.

The conditional mean of the observed categorical variable y in model (1) - (5) is given by (7). Accordingly, the mean of y conditional on x depends on x only through the index $x'\beta$. Therefore, (1)-(5) is a special case of the single index model given by

$$E\{y|x\} = H(x'\beta)\,, \tag{9}$$

where H is an unknown link function. If we relax the normality assumption (3) and replace it by the assumption

$$u \text{ is independent of } x\,, \tag{10}$$

we get the following expression instead of (7):

$$\begin{aligned} E\{y|x\} &= 2 - p_{2,1} + p_{2,3} - G(m_1 - x'\beta)(1 - p_{1,2} - p_{2,1} + p_{2,3} - 2p_{1,3}) + \\ &\quad [1 - G(m_2 - x'\beta)](1 - p_{3,2} - p_{2,3} + p_{2,1} - 2p_{3,1})\,, \end{aligned} \tag{11}$$

where G is the distribution function of the error term u ($G(t) = P[u \leq t]$).

Again, the right-hand side depends on x only through $x'\beta$, so that (1), (2), (4), (5) and (10) lead to the same single index model (9). The link function H is then given by G in (12). The crucial assumption here is that the misclassification probabilities in (4)- (5) do not depend on x. This is the typical identifying assumption in this type

of literature, used by Hausman et al. (1998), Lee and Porter (1984), but also in other applications such as Douglas et al. (1995). Such an assumption can only be avoided if a completely different measurement can be used as a benchmark, such as, in our empirical example, objective measurement of language proficiency (see Charette and Meng (1994)).

It is straightforward to extend (12) to the following results:

Proposition:

(a) If (1), (2), (5) and (6) are satisfied, then $E\{y|x\} = H(x'\beta)$ for some function H, i.e. the model is a single index model.

(b) If, moreover,

$$p_{1,2} + p_{2,1} - p_{2,3} + 2p_{1,3} \leq 1$$
$$\text{and} \tag{12}$$
$$p_{2,3} + p_{3,2} - p_{2,1} + 2p_{3,1} \leq 1,$$

then H can be chosen nondecreasing.

(c) If, moreover, (12) holds with strict inequalities and u has a continuous distribution with support (a, b), then H can be chosen strictly increasing and differentiable with positive derivative on the interval $(-b + m_1, -a + m_2)$.

We have shown that the models discussed above are all special cases of the general single index model (9) for some (unknown) link function H. In this model, the vector β of slope parameters is identified up to scale; the constant term is not identified. A number of estimators for β in this model have been discussed in the literature, under varying assumptions on the distribution of the explanatory variables x and regularity conditions on the link function H. Ichimura (1993) derives an asymptotically normal root n consistent estimator based upon nonlinear least squares combined with nonparametric estimation of H. This estimator has the drawback that it is computationally

burdensome, since it requires numerical minimization of a nonconvex objective function. Hausman et al. (1998) use the maximum rank correlation estimator of Han (1987). This also requires numerical optimization. Hausman et al. (1998) report that no convergence problems occurred in their Monte Carlo experiments or their empirical work. Our experience, however, is different: we ran into convergence problems, possibly due to the comparatively large number of explanatory variables.

Attractive from a computational point of view is the class of average derivative estimators or weighted average derivative estimators (see, for example, Powell et al. (1989)). These estimators require the distribution of x to be absolutely continuous, and are therefore not directly applicable to our empirical example. Horowitz and Haerdle (1996), however, have recently developed an estimator which builds upon a weighted average derivative estimator for the slope parameters (up to a scale normalization) of the continuous explanatory variables as a first step. The parameters of the discrete explanatory variables are estimated in a second step. Their estimator is consistent and asymptotically normal for all slope parameters (up to a normalizing scale parameter). Horowitz and Haerdle also show how to estimate the asymptotic covariance matrix consistently. The estimator does not require numerical optimization and is computationally very convenient. On the other hand, it requires the choice of a kernel (in the first step) and several smoothness parameters (in both steps). Procedures for choosing the smoothness parameters in the second step in an optimal way are not available, so that some ad hoc choices cannot be avoided. We will apply the Horowitz and Haerdle estimator and follow their choice of kernel and smoothness parameters.

One of the regularity conditions for the Horowitz and Haerdle estimator is a weak monotonicity condition on the link function H: H has to be monotonically increasing on some nonempty interval with a priori specified range. In the (parametric) model with misclassification probabilities, this condition is satisfied if the misclassification probabilities satisfy (12) with strict inequalities. Thus this regularity condition does not invalidate the claim that the model is more general than the parametric model

with explicit misclassification probabilities.

We briefly sketch the idea of the weighted average derivative estimator and the extension of Horowitz and Haerdle. Details can be found in the two papers referred to above. For continuous x, the weighted average derivative is given by

$$\delta = E\{f(x)\frac{\partial E\{y|x\}}{\partial x}\} = -2E\{y\frac{\partial f}{\partial x}\}\,. \tag{13}$$

Here $f(x)$ is the density of x. It can be estimated by a differentiable nonparametric kernel regression estimator $\hat{f}(x)$. The derivative of this function estimate is a nonparametric estimator of $\frac{\partial f}{\partial x}$. According to (13), an estimate of δ is obtained as -2 times the sample mean of the $y_i\frac{\partial \hat{f}(x_i)}{\partial x}$. Powell et al. (1989) show that this weighted average derivative estimator is consistent for δ, and derive its limit distribution (under appropriate regularity conditions).

In the single index model (9), we have

$$\delta = E\{f(x)G'(x'\beta)\}\beta\,. \tag{14}$$

Hence, the vectors δ and β are identical up to a scale factor. The weighted average derivative estimator can therefore be used to estimate β up to scale. By means of normalization, one of the slope coefficients is set to 1 or -1, so that the others are identified.

Now consider the case with both continuous and discrete regressors. Denote them by x and z, respectively, and write the single index model as

$$E\{y|x,z\} = H(x'\beta + z'\alpha)\,. \tag{15}$$

Horowitz and Haerdle (1996) partition the sample into subsamples with given values of z. Within each subsample, $z'\alpha$ is constant, and the model is a single index model in the continuous variables x only. This gives a consistent weighted average derivative estimator for β (which does not include the constant, and has one coefficient normalized to 1 or -1) for each subsample. Horowitz and Haerdle obtain a consistent but more

efficient estimator for β by combining these estimates, using minimum distance (i.e., they take the weighted average of the separate estimates, using the inverse of their estimated covariance matrices as weights).

To derive an estimator for α, let z_1 and z_2 be two different values of z corresponding to cells 1 and 2 in the partition, and let $H_1(x'\beta)$ and $H_2(x'\beta)$ be the within cell link functions. Thus $H_i(x'\beta) = H(x'\beta + z_i'\alpha)$, $i = 1, 2$. This gives a relation between H_1 and H_2 which is used by Horowitz and Haerdle (1996) to derive a condition which should be satisfied by α. Assume H is monotonically increasing on an interval with range $[c_0, c_1]$ (this is the weak monotonicity condition referred to above). Define $h_i(t) = max[c_0, min[H_i(t), c_1]]$ $(i = 1, 2)$. Then it is easy to see from a graph of h_1 and h_2, and straightforward to prove using some algebra, that

$$\int_{-\infty}^{\infty} [h_2(t) - h_1(t)]dt = (c_1 - c_0)(z_2 - z_1)'\alpha \tag{16}$$

This yields a moment restriction on α. Plugging in estimates of the link functions H_1 and H_2 yields an estimate of the left hand side of (16) and transfers (16) into a sample moment condition. Horowitz and Haerdle combine such sample moment conditions for different pairs z_1 and z_2, and thus derive a GMM type estimator for α. The estimator not only depends on the nonparametric estimators used for estimating β and the link functions, but also on the choice of c_0 and c_1. It is clear that these have to be chosen such that $[c_0, c_1]$ is contained in the range of H, but it is not clear what the optimal choice is, since the gains of using a larger interval $[c_0, c_1]$ and thus more observations, should be compared to the loss due to inaccurate estimation of the tails of H.

A final remark concerns the chosen cardinal scale of our observed dependent variable y, the outcomes of which we coded by 1, 2 and 3. If we change the coding to, e.g., 1, 2 and 4, this leads to a different link function, and to a different single index estimator. The link with the parametric model through the monotonicity condition also changes somewhat, since (12) will change. All single index estimators obtained with different codings will be consistent (under the appropriate assumptions, including the monotonicity condition), but it is not clear which one is most efficient. We do not

pursue this issue and only consider the coding 1, 2 and 3.

3 Application

3.1 Data and Variables

We apply both the parametric and the semiparametric estimator to data on speaking fluency of male immigrants in West–Germany. The data are drawn from the first (1984) wave of the German Socio-Economic Panel (GSOEP). We use the subsample of immigrant households from five typical guest worker countries in the 1950s and 1960s: Turkey, Yugoslavia, Italy, Greece and Spain. These households are oversampled in this wave of the GSOEP. We use only males who where older than 15 years at immigration. All survey questions are asked in the immigrant's home country language (see Dustmann (1994) for more details).

The dependent variable is a self–reported indicator of speaking fluency, reported on a five point scale. Due to the small number of observations in the extreme categories, we have transformed this information into a three level variable: $y_i = 3$ if individual i reports that he speaks German well or very well; $y_i = 2$ if he claims to speak the host country language on an intermediate level; $y_i = 1$ if his speaking fluency is bad or very bad.

The choice of explanatory variables is motivated by human capital theory. We use years since migration (YSM), age at entry (AGEENT), schooling (SCH), after school education (EDU), and dummy variables indicating the immigrant's nationality (T, Y, I, G) as regressors. The time of residence in the host country is a measure of exposure to the host country language, and we would expect individuals to improve their language fluency with the time in the host country. Age at entry is expected to affect language fluency negatively for two reasons: individuals who are older at entry may have a shorter pay off period for investments into language capital; and individuals' ability to

learn a new language may decrease with age. Individuals with higher levels of education should find it more easy to learn a new language, since higher education may reflect higher ability, and since education increases the productivity of accumulating language capital.[1]

We have also included country of origin dummies. Immigrants may be a self selected group. Since selection is determined by the economic conditions in home and host country, country of origin dummies may pick up level effects in the average ability level (see Borjas (1987)). Also, the relation between language proficiency and return migration may vary across the origin countries. Moreover, these dummies may reflect language distance and cultural differences, which affect the acquisition of language capital. Finally, origin dummies may capture enclave effects, if individuals from different origins have different propensities to live in ethnic communities.

Definitions and summary statistics of all the variables can be found in Table 1. The first four explanatory variables are measured in years and can be interpreted as continuous variables. The four dummy variables for nationalities, however, are obviously discrete.

3.2 Results

The estimation results are presented in Table 2. As explained above, one of the slope parameters has to be normalized to 1 or -1 for the semiparametric estimator. To make the parametric models comparable with the semiparametric model, we have used the same normalization in the parametric models. We have set the coefficient of AGEENT equal to -1. This variable has a significant negative effect and the largest absolute t-value if the parametric models are estimated with the usual normalization $\sigma = 1$.

[1]For instance, individuals who know how to read and to write learn a new language in a more systematic way than individuals who lack these skills. Also, the better educated may be more efficient in the acquisition of further knowledge. This reflects the idea that human capital is self productive in its own production (see Ben Porath (1967)).

Table 1: Variable Definitions and Sample Statistics

Variable	Mean	Std Dev	Minimum	Maximum
SPF	2.18	0.75	1	3
AGEENT	27.67	7.06	17	65
YSM	15.16	5.40	1	43
SCH	1.29	2.68	0	30
EDU	1.28	2.31	0	30
T	0.30	0.46	0	1
Y	0.21	0.41	0	1
G	0.14	0.35	0	1
I	0.21	0.41	0	1
S	0.14	0.35	0	1

SPF: speaking fluency (1: bad, 2: intermediate, 3: good)

YSM: years since migration to Germany

AGEENT: age at entry in Germany

SCH: years of schooling after the age of 14

EDU: years of job specific education after the age of 14

T,Y,G,I,S: dummies for nationalities:

Turkish (T), Yugoslavian (Y), Greek (G), Italian (I), Spanish (S)

Source: German Socio-Economic Panel 1984; 1185 observations

Table 2: Estimation Results

	Ordered Probit		Missclass. Model		Horowitz/ Haerdle	
	Coef	St er	Coef	St er	Coef	St er
constant	34.32	3.01	21.45	6.34		
T	-2.14	2.21	-1.66	2.10	-3.44	1.45
Y	13.12	2.75	12.28	2.63	7.25	2.30
G	5.33	2.64	6.65	2.47	3.85	3.87
I	4.71	2.40	5.48	2.33	1.16	1.74
YSM	0.35	0.13	0.43	0.13	0.29	0.10
AGEENT	-1.00	—	-1.00	—	-1.00	—
SCH	1.12	0.23	1.73	0.42	0.77	0.15
EDU	0.79	0.25	1.91	0.46	1.49	0.25
σ	19.95	1.89	10.34	4.67		
m_2	25.01	2.45	12.47	11.48		
$p_{1,2}$			0.225	0.124		
$p_{1,3}$			0.156	0.063		
$p_{2,1}$			0.069	0.267		
$p_{2,3}$			0.119	0.455		
$p_{3,1}$			0.036	0.023		
$p_{3,2}$			0.227	0.069		
Log-Likelihood	-1125.25		-1334.62			

The results for the ordered probit model are largely in accordance with other studies on language fluency. Years since migration, years of schooling and years of job specific education all have the expected positive effect on speaking fluency. The country dummies indicate that both the Spanish base group and Turkish workers have significantly lower probabilities to be fluent in German than the other groups.

Column 2 contains the estimates for the parametric model in which misclassification probabilities are explicitly included. Since these probabilities are by definition nonnegative, standard t-tests or likelihood ratio tests on $p_{k,j} = 0$ are inappropriate (see Shapiro (1985), for example). Still, the estimates of the $p_{k,j}$ and their standard errors imply that 0 is not contained in the one-sided 95% confidence intervals of $p_{1,2}$,

$p_{1,3}$ and $p_{3,2}$. This suggests that adding the probabilities of misclassification is indeed an improvement compared to the standard ordered probit model. While the estimates of $p_{1,2}$, $p_{1,3}$, $p_{3,1}$ and $p_{3,2}$ are rather precise, those of $p_{2,1}$ and $p_{2,3}$ have much larger standard errors, reflecting the problem that these are harder to identify. The estimated probabilities are small enough to satisfy the inequality conditions in (12). This implies monotonicity of the link function if the parametric model is written as a single index model, so that the monotonicity assumption required for the semiparametric estimator is fulfilled.

The qualitative effects of the regressors has not changed in this more general specification. However, some of the estimated slope coefficients in the second model differ substantially from those in the ordered probit model. In particular, the effect of the educational variables has increased considerably. The standard deviation of the error term u has decreased by almost 50 percent. This is because part of the unsystematic variation in observed speaking fluency is now explained by classification errors. The estimate of the bound m_2 has changed accordingly. In the next subsection, we will discuss what this implies for the estimated probabilities of bad, intermediate or good speaking fluency (i.e., the predictions of the model).

In column 3 of Table 2, the semiparametric estimates using the estimator of Horowitz and Haerdle (1996) are presented. The constant term is not estimated and, as before, the coefficient of AGEENT is normalized to -1. Note that the sign of this coefficient is identified, due to the assumption that the link function is increasing. We find the same sign as in the parametric models.

All the other coefficients also have the same sign as in the parametric models. The magnitudes of the eestimates change compared to the previous models. The effect of after school (job specific) education is stronger than in the ordered probit model, but the effect of general schooling is weaker. The effect of years of residence has decreased even further. The differences between semiparametric and parametric estimates of the coefficients on the home country dummies are larger than the differences between

the two parametric sets of estimates. Turkish immigrants are now significantly less fluent than the reference group of Spanish immigrants. Yugoslavian immigrants are still significantly more fluent than the Spanish, but the estimated difference is smaller. Greek and Italian immigrants are no longer significantly different from the Spanish immigrants.

The standard errors are not uniformly larger (and often even smaller) than in the parametric models. All estimators converge at the same rate, but if (one of) the parametric model(s) is not misspecified, the parametric ML estimator would be asymptotically more efficient. Smaller estimated standard errors for the semiparametric estimates can be due to finite sampling error in estimating the standard errors, or due to misspecification of the parametric models.

An obvious question arising with this type of semiparametric estimator is to which extent the results are sensitive to the choice of smoothness parameters. For the estimator at hand, this particularly applies to the choice of the smoothness parameters in the second estimation step (c_0 and c_1; see Section 2), which is the main novelty of the estimator. Rules for the choice of these smoothness parameters are not available. The results in the table are based upon $c_0 = 1.4$ and $c_1 = 2.6$. This choice corresponds to that of Horowitz of Haerdle, if the scale difference is accounted for (Horowitz and Haerdle use 0.2 and 0.8 for a variable which is zero or one, our dependent variable ranges from 1 to 3). We experimented with other choices. For example, for $c_0 = 1.8$ and $c_1 = 2.2$ we obtain (standard errors in parentheses): Turkish -4.54 (6.47); Yugoslavian 9.17 (6.26); Greek 4.68 (2.38), and Italian 2.67 (4.11). Thus the signs remain the same, but the estimates increase substantially in magnitude. Standard errors, however, increase even more, and all discrete variables become insignificant at the 5 percent level.

In figure 1, we have drawn the estimated link function H in (9).[2] The figure

[2] We use the quartic kernel. The bandwidth is chosen by visual inspection. This also holds for the nonparametric regressions in the next subsection.

also contains 95 percent uniform confidence bounds (based upon Haerdle and Linton (1994)). The estimated link function is increasing on its full domain, except at very low values of the index, for which the estimates are not very precise due to the small number of observations in that region. In an ordered response model without misclassification, the value of the link function should tend to 1 if the index value tends to $-\infty$. The figure suggests that this is not the case. This could be due to misclassification of those with low speaking fluency ($y = 1$).

3.3 Comparison of the Three Models

In this subsection we compare the three models. First, we look at their predictions, i.e. the estimated (conditional) probabilities of bad, intermediate and good speaking fluency (given x), or the conditional mean of the outcome y or z coded 1, 2 or 3, which is a linear combination of these three probabilities. In the ordered probit model, observed and true speaking fluency (y and z) coincide, but in the model with misclassification they do not. Comparing predictions of observed and true speaking fluency should tell us how different the implications of the two parametric models are. The semiparametric model only identifies the observed speaking fluency probabilities, and we compare these with those of the two parametric models. Finally, we formally test for misspecification of the parametric models, using a graphical test against a semiparametric single index alternative.

For the parametric models, the predictions are straightforward functions of the estimated parameters. For given parameter values, they are completely determined by the model specification. For the semiparametric model, however, this is not the case. Predictions can be obtained by nonparametric regression of (a function of) the dependent variable on the estimated index $x'b$. Nonparametric regressions of dummies for good (including very good), bad (including very bad) or intermediate speaking fluency on the index, yield predicted probabilities of good, bad or intermediate fluency. Nonparametric regression of y yields a prediction of $E\{y|x\}$. The latter nonparametric

Table 3: Predictions of Parametric and Semiparametric Models

sample means and (in ()) sample standard deviations

	Ordered Probit	Misclassification Model		Horowitz/ Haerdle
		reported (y)	latent (z)	
$\hat{E}\{y_i\|x_i\}$ or $\hat{E}\{z_i\|x_i\}$	2.175	2.177	2.136	2.176
	(0.308)	(0.321)	(0.547)	(0.315)
$\hat{P}\{y_i = 1\|x_i\}$ or $\hat{P}\{z_i = 1\|x_i\}$	0.208	0.208	0.277	0.207
	(0.138)	(0.154)	(0.264)	(0.153)
$\hat{P}\{y_i = 3\|x_i\}$ or $\hat{P}\{z_i = 3\|x_i\}$	0.383	0.384	0.413	0.383
	(0.175)	(0.175)	(0.297)	(0.176)

regression is the same as the nonparametric estimator of H, given in Figure 1 above.

In Table 3, some summary statistics of the predictions are presented. The means and standard deviations of the predictions of the observed outcomes are similar for the three models. The mean predictions are also similar to the sample means of the outcomes. Correlation coefficients (not presented in the table) also appear to be quite large, ranging from 0.90 to 0.97.

Larger differences are found with the predictions of the true outcomes according to the misclassification model. In particular, the average predictions of bad or good true fluency are larger than the corresponding predictions for observed fluency. Accordingly, the sample dispersion of the predictions for z, the speaking fluency variable free of misclassification error, is larger than that for the predictions of y, the observed speaking fluency indicator.

In Figure 2, we present a scatter plot of the predicted probabilities of speaking the language well or very well according to the two parametric models. For the misclassification model (vertical axis), Figure 2 shows the predictions of the latent variable z. For the ordered probit model (horizontal axis and 45 degree line), predictions of y and z coincide. We find that the misclassification model leads to more probability estimates

close to zero or one than the ordered probit model, explaining the large dispersion in $\hat{P}[z = 3|x]$ according to the misclassification model. Still, the correlation between the two sets of predictions is quite large (the sample correlation coefficient is 0.96).

This is very different in Figure 3, where we compare predictions of the probability that individuals *report* good or very good speaking fluency. In the misclassification model, due to the possibility of reporting errors, even for those with very high or very low probability of actually being fluent, the probability that they report being fluent is not close to one or zero. For observations with less extreme predictions, the predictions according to ordered probit and misclassification models are similar, with some exceptions. Again, the correlation coefficient is about 0.96.

The substantial differences between latent and observed outcomes in the misclassification model confirm the conclusion from the misclassification probabilities in Table 2: generalizing the ordered probit model by incorporating misclassification probabilities is useful in this empirical example. The high correlation coefficients reflect the similarity in the ordered probit and misclassification model estimates of β, leading to similar estimates for $x'\beta$. Predictions for y and z are different functions of this index. While the predictions for the reported variable y are similar for ordered probit and misclassification model, except for observations in the tails, the predictions for the latent variable z are not.[3]

In the next two figures, we compare the predictions of the observed index value $E\{y|x\}$ according to the semiparametric model with those of the ordered probit model (Figure 4), and those of the misclassification model (Figure 5).[4] Most points in the scatter plot are near the 45 degree line, indicating that predictions are generally similar.[5] There are some exceptions, however. In particular, ordered probit leads

[3]We come to the same conclusions when we draw figures of the probability of bad or very bad speaking fluency, or of the expected value of y or z. These figures are not reported.

[4]Similar conclusions are obtained from figures comparing parametric and semiparametric predictions of $P[y = 1|x]$ or $P[y = 3|x]$. We do not report these.

[5]This is confirmed by the sample correlation coefficients of the predictions: 0.95 between semipara-

to more predictions at the lower and at the upper end of the interval $[1,3]$. For the misclassification model, predictions are never larger than 2.71 or smaller than 1.54, due to the misclassification probabilities.[6] The range of predictions for the semiparametric model is about the same, if we ignore a few outliers in the tails, for which the semiparametric predictions are very inaccurate.

Misspecification Tests of Parametric Models

In Figure 6 and 7, we present graphical tests of the two parametric models against the semiparametric single index alternative. These tests are similar to those proposed by Horowitz (1993) for the parametric binary choice model. The null hypothesis is that the parametric model is correctly specified (ordered probit in Figure 6, Misclassification model in Figure 7). The alternative is that the parametric model is not correctly specified; the test should have some power in the direction of the semiparametric alternative which we discussed, but it is not clear whether it has power in other directions of misspecification (such as other than single index models).

Each figure presents two functions of the index estimate $x'b/s$, where b and s are the parametric estimates of β and σ in Table 2 (ordered probit estimates in Figure 6, misclassification model estimates in Figure 7; the semiparametric estimates in Table 2 are not used). The solid line reflects the predicted probabilities $\hat{P}[y_i = 3|x_i] = \hat{P}[y_i = 3|x_i'b/s]$ according to the parametric model, as a function of $x_i'b/s$. The circles are nonparametric kernel regression estimates of the observed dummy indicator variable $I(y_i = 3)$ on the same index $x_i'b/s$. The dashed lines are nonparametric uniform 95% confidence bands around these kernel estimates.[7]

---

metric and ordered probit predictions of $E\{y|x\}$, 0.96 between semiparametric and misclassification model estimates of $E\{y|x\}$.

[6] For $z \to -\infty, H(z) \to 1 + p_{1,2} + 2p_{1,3}$, which equals -1.54 according to the estimates in Table 2. A similar argument applies for $z \to \infty$.

[7] Since the estimator b/s converges to β/σ at rate root n, which is faster rate than the rate of convergence of the nonparametric estimator, the standard errors of b and s are asymptotically negligible,

Under the null hypothesis that the parametric model is specified correctly, then b/s is consistent for β/σ. In that case, the parametric formula for the predicted probability $\hat{P}[y_i = 3|x_i]$ is consistent for $P[y_i = 3|x_i]$. The null hypothesis, however, also implies that $P[y_i = 3|x_i]$ is a single index function of $x_i'\beta$, and b/s is a consistent estimate of this single index (up to scale). The nonparametric (circled) curve is the corresponding estimated link function, and it will also be consistent for $P[y_i = 3|x_i]$. Thus, under the null, both curves are consistent for the same function, and should be similar. A test of the null hypothesis can be performed by testing whether the circled curve is significantly different from the solid curve. Since the solid curve is based upon parametric estimates which converge at rate $\sqrt{n}$, while the circled curve converges at the lower rate $n^{0.4}$, the imprecision in the solid curve can be neglected compared to that in the circled curve, and an asymptotically valid test can be based upon the uniform confidence bands around the circled curve. In both figures, the solid curve is everywhere between the uniform confidence bands. Thus in neither case, the parametric model can be rejected.[8]

That we cannot reject the parametric model is somewhat surprising in Figure 6, since we already concluded from Table 2 that the ordered probit model fits the data much worse than the parametric misclassification model. An explanation could be lack of power of the test. In particular, most of the difference between ordered probit and misclassification model predictions is for values of the index in the tails (see Figure 3), where the nonparametric estimates are not very accurate due to lack of observations. This is reflected by the large distance between upper and lower confidence bounds in Figure 6 at the lower and upper end. In general, little can be said about the power of this type of tests. Under the alternative that the parametric model is misspecified, estimates of β/σ will typically be inconsistent, and neither the parametric nor the semiparametric estimates of $P[y_i = 3|x_i]$ will be consistent. What this implies for the difference between them, however, is not clear.

---

and the uniform confidence bands are calculated as if b/s were known.

[8]The same conclusion is obtained if $P[y_i = 1|x_i'b]$ or $E\{y_i|x_i'b\}$ are used instead of $P[y_i = 3|x_i]$. These figures are not reported.

4 Summary and Conclusions

In models with ordered categorical dependent variables where the categorical assignment is based on subjective self-reported evaluations, misclassification is likely to be considerable, and may lead to seriously biased parameter estimates, as well as biased predictions. Parametric estimators which incorporate and estimate misclassification probabilities, as well as semiparametric estimators, are an alternative to standard parametric models. Extending the work of Lee and Porter (1984) and Hausman et al. (1998), we introduce a parametric model which incorporates misclassification probabilities for the case of more than two ordered categories. We show that this model is a special case of a semiparametric single index model which, if misclassification probabilities are not too large, satisfies some monotonicity condition. Therefore, it can be estimated with a recently developed estimator of Horowitz and Haerdle (1996).

We analyse the determinants of immigrants' language proficiency, and compare the results of the standard model with those of the parametric model with misclassification and with the semiparametric results. In all models, the signs of the estimated slope coefficients are the same. Magnitudes and significance levels of the effects vary, however.

We find that the parametric misclassification model is a significant improvement compared to ordered probit. Some of the estimated misclassification probabilities are substantially larger than zero, and incorporating them leads to a much better fit of the data. In the misclassification model, the predictions of true speaking fluency deviate substantially from those of reported fluency. Predicted probabilities of reported fluency are similar for both parametric models and for the semiparametric models, except for those observations with a very small or very large value of the underlying index. A formal test of the parametric models versus a semiparametric alternative does not reject either of the parametric models.

When analysing categorical variables which are likely to suffer from misclassification, the misclassification model and the semiparametric estimator we have suggested appear to be a substantial improvement. The parametric misclassification model is

also easy to implement, and it gives predictions of the true categorisation. It also provides estimates of the mislassification probabilities, which may be of their own interest. A shortcoming of the model is that probabilities of misclassification in intermediate ctegories are not precisely estimated, since their identification relies on paramteric assumptions. Better estimates of all misclassification probabilities would require additional data, for example alternative measurements (Charette and Meng (1994)), or panel data. This is on our research agenda.

References

- Ben-Porath, Y. (1967): The Production of Human Capital and the Life Cycle of Earnings, *Journal of Political Economy*, 75, 352-365
- Borjas, G.J. (1987): Self-Selection and the Earnings of Immigrants, *American Economic Review*, 77, 531-553
- Charette, M. and R. Meng (1994), Explaining Language Proficiency, *Economics Letters*, 44, 313-321.
- Chiswick, B. (1991), Reading, speaking, and earnings among low-skilled immigrants, *Journal of Labor Economics*, 9, 149-170.
- Chiswick, B. and P. Miller (1995), The Endogeneity between Language and Earnings: International Analyses, *Journal of Labor Economics*, 13, 246-288.
- Clark, A. and A. Oswald (1994), Unhappiness and unemployment, *Economic Journal*, 104, 648-659.
- Das, M. and A. van Soest (1997), Expected and realized income changes: Evidence from the Dutch socio-economic panel, *Journal of Economic Behavior and Organization*, 32, 137-154.

- Douglas, S., K. Smith Conway and G. Ferrier (1995), A switching frontier model for imperfect sample separation information: with an application to labor supply, *International Economic Review*, 36, 503-527.

- Dustmann, C. (1994), Speaking fluency, writing fluency and earnings of migrants, *Journal of Population Economics*, 7, 133-156.

- Dustmann, C. and A. van Soest (1998), Language and the earnings of immigrants, CEPR discussion paper series No. 2012.

- Haerdle, W. and O. Linton (1994), Applied nonparametric methods, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume IV, North-Holland, Amsterdam.

- Han, A.K. (1987), Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator, *Journal of Econometrics*, 35, 303-316.

- Hausman, J., J. Abrevaya and F. Scott-Morton (1998), Misclassification of a dependent variable in a discrete response setting, *Journal of Econometrics*, 87, 239-269.

- Horowitz, J. (1993), Semiparametric estimation of a work-trip mode choice model, *Journal of Econometrics*, 58, 49-70.

- Horowitz, J. and W. Haerdle (1996), Direct semiparametric estimation of single index models with discrete covariates, *Journal of the American Statistical Association*, 91, 1632-1640.

- Ichimura, H. (1993), Semiparametric least squares (SLS) and weighted SLS estimation of single index models, *Journal of Econometrics*, 58, 71-120.

- Kerkhofs, M. and M. Lindeboom (1995), Subjective health measures and state dependent reporting errors, *Health Economics*, 4, 221-235.

- Lee, L.F. and R.H. Porter (1984), Switching regression models with imperfect sample information with an application on cartel stability, *Econometrica*, 52, 391-418.

- McManus, W., W. Gould, and F. Welch (1983), "Earnings of Hispanic men: the role of English language proficiency", *Journal of Labor Economics*, 1, 101-130.

- Powell, J., J. Stock, and T. Stoker (1989), Semiparametric estimation of index coefficients, *Econometrica*, 57, 1403-1430.

- Rivera-Batiz, F. (1990), "English language proficiency and the economic progress of immigrants", *Economics Letters* , 34, 295-300.

- Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrica*, 72, 133-144.

Figure 1. Link Function Semiparametric Model

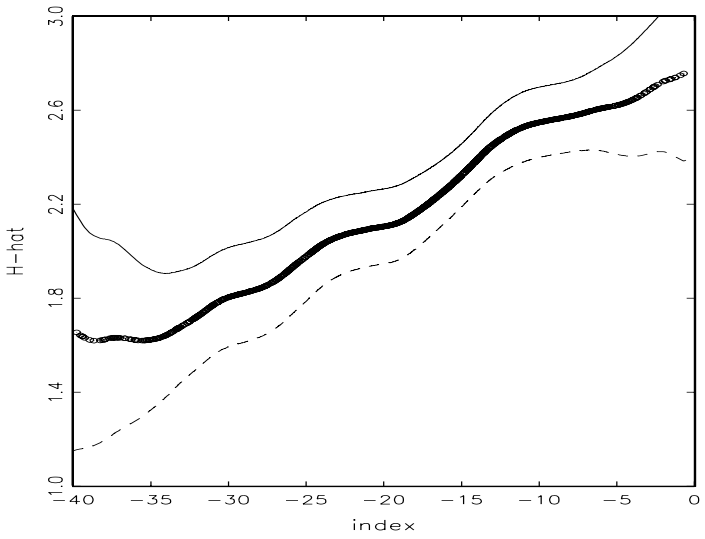

Figure 2. P(good or very good); Parametric Models

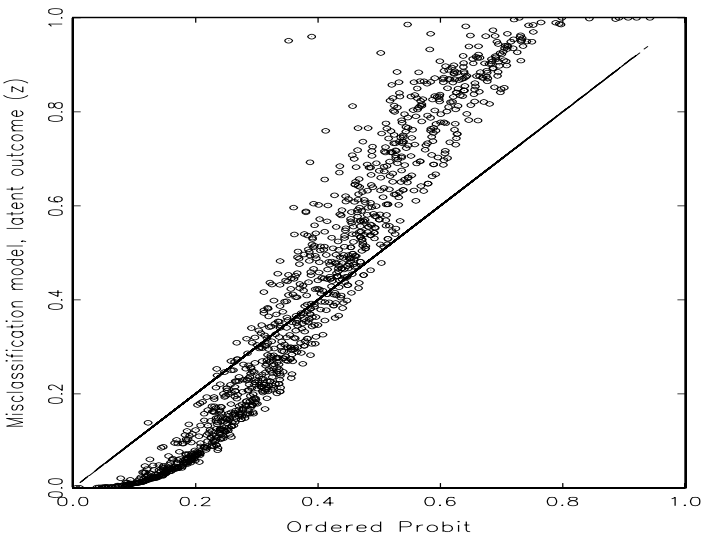

Figure 3. P(good or very good): Parametric Models

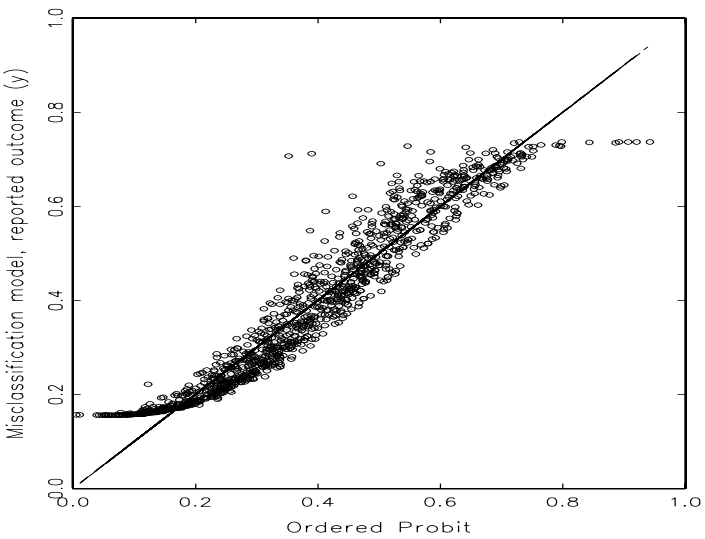

Figure 4. Expected value of reported outcome (y)

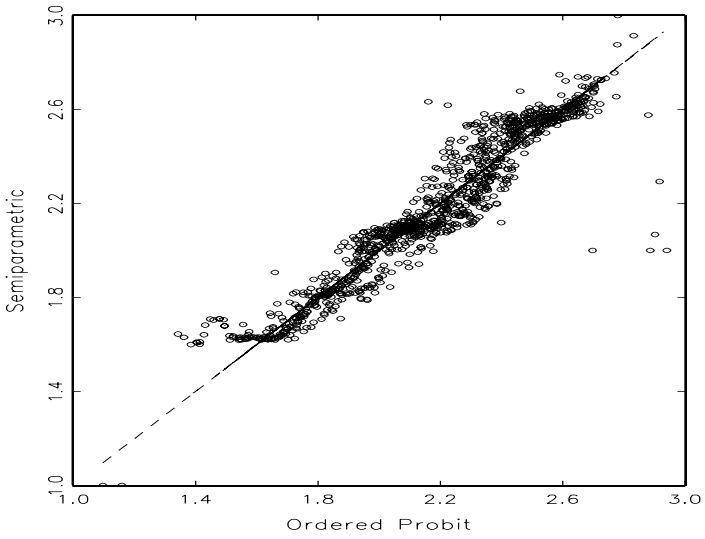

Figure 5. Expected value of reported outcome (y)

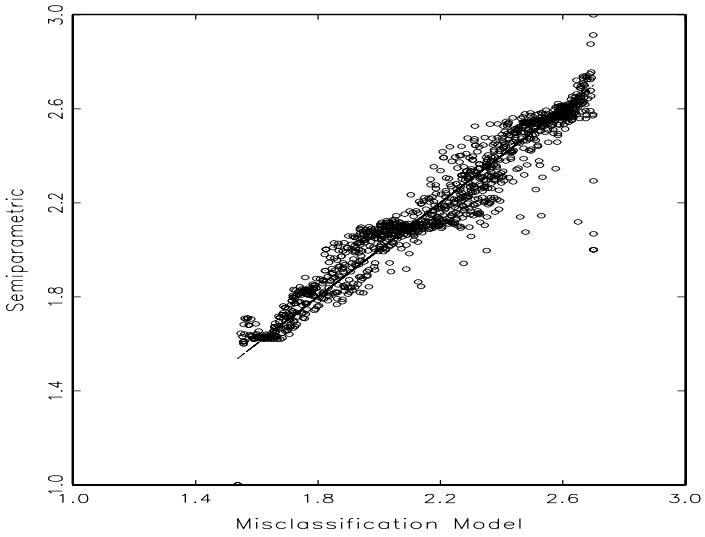

Figure 6. Test of Ordered Probit Model

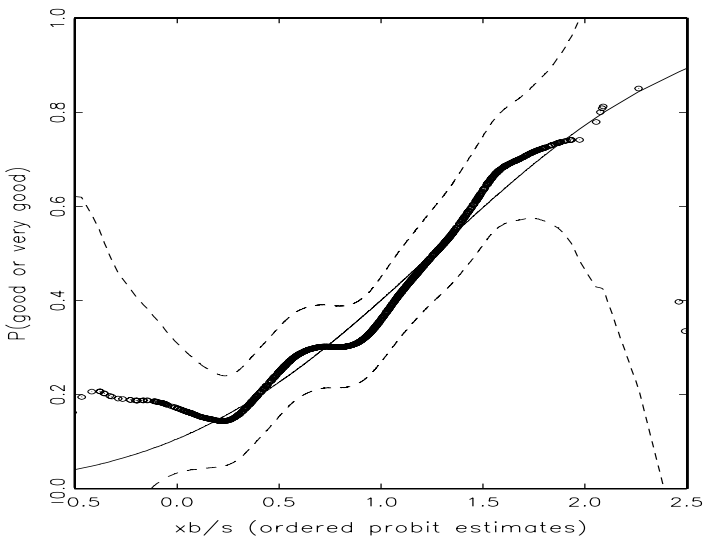

Figure 7. Test of Parametric Misclassification Model